DEPARTMENT OF STATISTICS

THE OHIO STATE UNIVERSITY

COLUMBUS, OHIO

6

APPROXIMATE MAXIMUM LIKELIHOOD ESTIMATES IN REGRESSION MODELS FOR GROUPED DATA

by

10 A. Indrayan J. S. Rustagi
Department of Preventive and Social Medicine
University of Gorakhpur, Gorakhpur, India

and

J. S. Rustagi*
Department of Statistics
The Ohio State University
Columbus, Ohio

9 Technical rept.

11 1978

12 33 p.

14 TR-165

Tech. Report No. 165
1978

*Supported in part by Contract 15 N00014-78-C-0543 (NR 042 403) by Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

INTRODUCTION

Grouped data arise quite frequently in application. Various experimental situations naturally lead to observations classified into certain groups. The process of recording or storing observations directly leads to grouped data. Other examples of circumstances which lead to grouped data have recently been discussed by Haitovsky (1973) in his mongraph on regression estimation from grouped observations. From a practical point of view, the case for grouped data has been advocated by Durbin (1954) as a technique for minimizing errors of measurements. Grouped observations also result in experiments where precise measuring instruments are not available.

There is an extensive literature on the study of grouped data. The estimates of probability density function as a histogram, by the very nature of the problem require grouped observations. Standard tests of goodness of fit require grouping of data for performing tests such as those using the Chi squared statistic. The effect of grouping in regression problems has been studied by many authors and Haitovsky has provided a large list of references. The computational problems involved in obtaining exact solutions become difficult. In this paper, approximate estimates of the parameters in the regression models have been proposed. Certain optimal properties of these estimates are discussed and a comparison of these estimates with other competing ones is given. Simulation experiments show that the approximate estimates based on grouped data compare well with those based on ungrouped data under certain cases.

2. MODEL AND NOTATION

We consider K populations generated by the random variables Y_1, $Y_2, \ldots,$ Y_K. Suppose the possible values of the random variables occur in the disjoint intervals (c_o, c_1) $(c_1, c_2) \ldots (c_{I-1}, c_I)$. and only the frequencies of the random variables are known in these intervals. We use the following notation:

(a) N_{ik} = number of Y_k's in the
interval (c_{i-1}, c_i),
$$i = 1, 2, \ldots, I, \; k = 1, 2, \ldots, K. \tag{2.1}$$

(b) ϕ be the standard normal density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \tag{2.2}$$

$$\Phi(x) = \int_{-\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt. \tag{2.3}$$

Let Y_k be normally distributed with mean μ_k and variance σ^2 with

(c) $\mu_k = \beta_1 x_{k1} + \beta_2 x_{k2} + \ldots + \beta_p x_{kp}$,
$$k = 1, 2, \ldots, K. \tag{2.4}$$

(d) $$\zeta_{ik} = \frac{C_i - \mu_k}{\sigma}, \tag{2.5}$$

(e) $$\pi_{ik} = \Phi(\zeta_{ik}) - \Phi(\zeta_{ik-1}) \tag{2.6}$$

The likelihood of the sample of size n_1, $n_2, \ldots, n_K$ from the K populations, is the product of multinomials, given by

$$L(\underset{\sim}{\theta}) = \prod_{k=1}^{K} \left(\frac{n_k!}{N_{1k}! N_{2k}! \ldots N_{Ik}!} \pi_{1k}^{N_{1k}} \ldots \pi_{Ik}^{N_{Ik}} \right) \tag{2.7}$$

with $n_k = N_{1k} + N_{2k} + \ldots + N_{Ik}$

$$n = \sum_{k=1}^{K} n_k \; ;$$

and $\underset{\sim}{\theta} = (\beta_1 \;\; \beta_2 \; \cdots \; \beta_p \;\; \sigma^2)'$ (2.8)

is the (p+1) dimensional vector of parameters.

The likelihood equations for estimating $\underset{\sim}{\theta}$ are obtained by differentiating (2.7) with respect to $\beta_1, \beta_2, \ldots, \beta_p$ and σ^2. We then have, using (2.6)

$$\sum_i \sum_k x_{k\ell} \frac{N_{ik}[\phi(\zeta_{ik}) - \phi(\zeta_{i-1\ k})]}{\Phi(\zeta_{ik}) - \Phi(\zeta_{i-1\ k})} = 0$$

$$\ell = 1, 2, \ldots, p \tag{2.9}$$

and

$$\sum_i \sum_k N_{ik} \frac{\phi'(\zeta_{ik}) - \phi'(\zeta_{i-1\ k})}{\Phi(\zeta_{ik}) - \Phi(\zeta_{i-1\ k})} = 0 \tag{2.10}$$

The form of the above likelihood equations makes the estimation of the parameters an involved problem. In the case, p=1, Gjeddeback (1949) has derived the likelihood equations (2.9) and (2.10). By introducing certain tabulated functions, he was able to solve the equations in case p=1. Later Kulldorff (1961) gave conditions under which the maximum likelihood estimates exist and are unique roots of the likelihood equations for p=1, for an estimate of β_1 as well as for an estimate of σ^2 separately. Simultaneous estimation of β_1 and σ^2 was not attempted.

Using approximations to π_{ik}, an alternative approach is given here to find maximum likelihood estimates of the parameters in the case of the multiple regression model. Several questions concerning the uniqueness of solutions and the convergence of the solutions obtained by iterative procedures are still under study. The desirability of using approximate estimates arises from the fact that they have certain nice properties, some of which are discussed here.

3. APPROXIMATE MAXIMUM LIKELIHOOD ESTIMATES

The general approach here is to assume that certain terms in the expansion of the probabilities Π_{ik} are negligible. For this purpose, we assume that the intervals (c_{i-1}, c_i), $i = 1, 2, \ldots, I$ are of the same length h. That is

$$c_i - c_{i-1} = h,\ i = 1, 2, \ldots, I \tag{3.1}$$

We assume that the numbers of intervals, I, is infinite. Let m_i be the middle point of the interval (c_{i-1}, c_i), such that the random variable M may be defined with

$$P(M = m_i) = \Pi_{ik}, \tag{3.2}$$

and let $\xi_{ik} = (m_i - \mu_k)/\sigma$, $i = 1, 2, \ldots, I$. Note that $M = {}_{mi}$ if and (3.3) only if $c_{i-1} < y < c_i$. We have,

$$\Pi_{ik} = \Phi(\xi_{ik} + \frac{h}{2\sigma}) - \Phi(\xi_{ik} - \frac{h}{2\sigma})$$

$$= \int_{\xi_{ik} - \frac{h}{2\sigma}}^{\xi_{ik} + \frac{h}{2\sigma}} \sum_{r=0}^{\infty} \frac{(z - \xi_{ik})^r}{r!} \phi^{(r)}(\xi_{ik})\, dz$$

$$= \frac{h}{\sigma} \phi(\xi_{ik}) \sum_{r=0}^{\infty} \frac{(\frac{h}{\sigma})^{2r}}{2^{2r}(2r+1)!} H_{2r}(\xi_{ik}) \tag{3.4}$$

where $H_{2r}(\xi_{ik})$ is a Hermite polynomial of order 2r. The detailed study of Hermite polynomials has been given by Kendall and Stuart (1969 p. 155-156). Using Taylor's expansion, we have

$$\log \pi_{ik} = \log \frac{h}{\sigma} - \log \sqrt{2\pi} - \frac{1}{2} \xi^2_{ik} + \frac{h^2}{24\sigma^2} (\xi^2_{ik} - 1) + \frac{h^4}{2880\sigma^4} (\xi^4_{ik} + 4\xi^2_{ik} - 2) + 0\left(\left(\frac{h}{\sigma}\right)^4\right). \tag{3.5}$$

Estimation of $\underset{\sim}{\beta}$

The likelihood equations are

$$\sum_k \sum_i N_{ik} \frac{\partial \log \pi_{ik}}{\partial \beta_\ell} = 0, \qquad \ell = 1, 2, \ldots, p. \tag{3.6}$$

Using (3.5), the equation (3.6) reduces to

$$\sum_k \sum_i N_{ik} \left(- \frac{x_{k\ell}}{\sigma}\right) \left[-\xi_{ik} + \frac{h^2}{12\sigma^2} \xi_{ik} - \frac{h^4}{720\sigma^4} \left(\xi^3_{ik} + 2\xi_{ik}\right)\right] + 0_\ell\left(\frac{h^4}{\sigma^4}\right) = 0 \tag{3.7}$$

Substituting expressions for ξ_{ik} from (3.3) and μ_k from (2.4), we have on simplification, the likelihood equations (3.7) reduce to

$$\beta_1 \sum_i \sum_k N_{ik} x_{k\ell} x_{k1} + \beta_2 \sum_i \sum_k N_{ik} x_{k\ell} x_{k2} + \ldots + \beta_p \sum_i \sum_k N_{ik} x_{k\ell} x_{k2} = \sum_i \sum_k N_{ik} m_i x_{k\ell} + 0_\ell \left(\frac{h^4}{\sigma^4}\right). \tag{3.8}$$

Ignoring terms $0_{\ell}(\frac{h^4}{\sigma^4})$, the likelihood equations are standard equations in the form

$$\hat{\underset{\sim}{\beta}}_o = (\underset{\sim}{x}'\underset{\sim}{x})^{-1}\,\underset{\sim}{x}'\,\underset{\sim}{m} \qquad (3.9)$$

where $\underset{\sim}{m}$ is the vector of mid points of the intervals and $\underset{\sim}{x}$ is the matrix of independent variables $x_{k\ell}$. $\underset{\sim}{x}$ is assumed to be of full rank.

Estimation of σ^2 with the above approximation is obtained by the equation

$$1 - \frac{1}{n}\sum_{i}\sum_{k} N_{ik}\,\xi_{ik}^2 + \frac{h^2}{12\sigma^2}\left(\frac{2}{n}\sum_{i}\sum_{k} N_{ik}\,\xi_{ik}^2 - 1\right) = 0 \qquad (3.10)$$

Let μ_k be estimated by

$$\hat{\mu}_{ok} = \frac{\sum_{i} N_{ik}\,m_i}{n_k}\,, \qquad (3.11)$$

and let us denote by

$$s_{02} = \frac{1}{n}\sum_{i}\sum_{k} N_{ik}\,(m_i - \hat{\mu}_{ok})^2\,, \qquad (3.12)$$

The equation (3.10) reduces to

$$1 - \frac{s_{02}}{\sigma^2} + \frac{h^2}{12\sigma^2}\left(2\,\frac{s_{02}}{\sigma^2} - 1\right) = 0 \qquad (3.13)$$

leading to an estimate (within $0(\frac{h^4}{\sigma^4})$)

$$\hat{\sigma}^2 = s_{02}\left[1 - \frac{h^2}{12s_{02}}\right] \qquad (3.14)$$

There are other methods of obtaining approximate estimates of $\underset{\sim}{\beta}$ and σ^2, given by Tallis (1967) and Fryer and Pathybridge (1972). These authors have used iterative procedures, arriving at the same results as (3.9) and (3.14). In the calculations of these estimates by the above method, iteration is avoided and consequently, there is considerable amount of saving.

With the above approximation, that is, within the terms of $O(\frac{h^4}{\sigma^4})$, the estimates correspond to ordinary least squares estimates. Such estimates are known to have many desirable properties.

We also note that to the order of approximation implied in Shappard corrections,

$$E(\hat{\underset{\sim}{\beta}}_o) = \underset{\sim}{\beta} \qquad (3.15)$$

and

$$\text{Cov}\ (\hat{\underset{\sim}{\beta}}_o) = (\underset{\sim}{x}'\underset{\sim}{x})^{-1}\ \sigma^2\ (1 + \frac{h^2}{12\sigma^2})\ . \qquad (3.16)$$

Sheppard corrections also lead to the estimates (3.16) when it is assumed that h is small. Here we are assuming that h/σ is small. The estimates obtained by the above method eliminate the need of iteration and can be computed by available computer programs.

The loss of efficiency in estimating $\underset{\sim}{\beta}$ by the mid point estimator $\hat{\underset{\sim}{\beta}}_o$ as compared to the regular ungrouped estimator $\hat{\underset{\sim}{\beta}}$ is approximately $\frac{1}{12} \cdot \frac{h^2}{\sigma^2}$. By a proper choice of h, therefore, the efficiency loss can be made as small as one would like. In mortality studies, where grouped data are usually obtained, the loss of efficiency can be compared with the cost of experiment. For some further interesting results, the reader may refer to Thompson (1977).

4. SIMULATION RESULTS

Comparison of approximate estimates with the actual estimates is not easy since the actual estimates under grouping require tedious computations. We shall compare here the estimates for certain simulated experiments under the assumptions of ungrouped and grouped observations.

Two experiments were performed. The first experiment is concerned with the following model. It is assumed that y_{ik} are normally distributed with the same variance σ^2, and

$$E(y_{ik}) = \alpha \ + \beta \ x_k$$

$$k = 1, 2, \ldots, K. \quad i = 1, 2, \ldots, n_K.$$

Only a selected values of the parameters were considered. Normal random deviates were chosen for each possible combination of parameters and 100 replications of the samples were obtained. The following parameter values were selected as given in Table 1.

Table 1

	α_0	α_1	σ^2
(i)	10	0	25
(ii)	40	0	25
(iii)	10	1	100
(iv)	40	0	100

for h = 0, 2, 3, 4, 5, 10, 15, 20

Also the number of observations chosen were $n_k = 10$ and 100 and $K = 10$. $x_1 = 0$, $x_2 = 1, \ldots, x_{10} = 9$. The results are given in tables 2-5.

The second model considered is the following.

$$y_{ik} = \beta_1 x_{k1} + \beta_2 x_{k2} + \beta_3 x_{k3} \,,$$
$$i = 1, 2, \ldots, n_k, \; k = 1, 2, \ldots, K \,.$$

We consider the cases for

$\beta_1 = 0$,

$\beta_2 = 0$,

$\beta_3 = 0$,

$\sigma^2 = 10{,}000$ and

$n_k = 10, 25, 100$,

$K = 4$.

The 3xn matrix is chosen to be

$$\underset{\sim}{X}' = \begin{pmatrix} 0 & 0 & \ldots & 0 & 2 & 2 & \ldots & 2 & 7 & 7 & \ldots & 7 & 9 & 9 & \ldots & 9 \\ 1 & 1 & \ldots & 1 & 4 & 4 & \ldots & 4 & 5 & 5 & \ldots & 5 & 1 & 1 & \ldots & 1 \\ 3 & 3 & \ldots & 3 & 6 & 6 & \ldots & 6 & 3 & 3 & \ldots & 3 & 8 & 8 & \ldots & 8 \end{pmatrix}$$

$h = 0, 10, 50, 100, 200$.

Tables 6-9 provide the estimates of the parameters $\underset{\sim}{\beta}$ and σ^2. The ungrouped observations are obtained when $h = 0$. The tables show clearly that mid point estimates correspond very well with the ungrouped estimates.

5. APPLICATION

The theory developed above has been applied to a study recently completed by Student Health Services of the Ohio State University. The data were collected in 1975 to study blood pressure levels and their correlates for the patients visiting the Student Health Services. In addition to Systolic and Diastolic blood pressures, other physiological variables

Table 2: Mean and S.E. of ungrouped and mid-point esimates of α, obtained in the simulation-model (i), $n_k=10$

h	Mean	S.E.	$(S.E.)^1/(1+h^2/12\sigma^2))$
$\sigma^2=25$ $\alpha=10$			
0	9.9565	1.0006	1.0006
2	9.9528	1.0120	1.0073
3	9.9691	1.0401	1.0155
4	9.9388	1.0094	1.0270
5	9.9344	1.0454	1.0415
10	9.9705	1.1260	1.1554
15	9.4664	1.2471	1.3237
20	9.9618	0.7274	1.5285
$\sigma^2=25$ $\alpha=40$			
0	39.9387	0.9144	0.9144
2	39.9409	0.9330	0.9204
3	39.9563	0.9238	0.9280
4	39.9081	0.9475	0.9384
5	39.8800	0.9328	0.9517
10	39.8169	0.9069	1.0558
15	39.5888	1.3240	1.2150
20	39.6673	1.6212	1.3967
$\sigma^2=100$ $\alpha=10$			
0	10.1800	1.9060	1.9060
2	10.1743	1.9352	1.9092
3	10.1605	1.9118	1.9131
4	10.2079	1.9485	1.9187
5	10.1628	1.9267	1.9258
10	10.1169	2.0492	1.9838
15	10.1877	2.0107	2.0770
20	10.2509	2.3219	2.2009
$\sigma^2=100$ $\alpha=40$			
0	40.1767	2.0242	2.0242
2	40.1545	2.0245	2.0276
3	40.1892	2.0221	2.0318
4	40.1753	2.0289	2.0377
5	40.1889	1.9825	2.0452
10	40.2031	2.0785	2.1069
15	40.1250	2.2190	2.2059
20	40.3545	2.3866	2.3374

[1] (S.E.) for ungrouped data, i.e., for h=0

Table 3: Mean and S.E. of ungrouped and mid-point estimates of β, obtained in the simulation-model (i), $n_k=10$

h	Mean	S.E.	$(\text{S.E.})^1\sqrt{(1+h^2/(12\sigma^2))}$
		$\sigma^2=25$ $\alpha=10$	
0	0.0063	0.1951	0.1951
2	0.0061	0.1971	0.1964
3	0.0037	0.1994	0.1980
4	0.0100	0.1942	0.2002
5	0.0034	0.2072	0.2030
10	-0.0015	0.2229	0.2253
15	-0.0056	0.2453	0.2581
20	0.0018	0.1506	0.2980
		$\sigma^2=25$ $\alpha=40$	
0	0.0124	0.1532	0.1532
2	0.0125	0.1598	0.1542
3	0.0109	0.1553	0.1555
4	0.0150	0.1631	0.1572
5	0.0193	0.1608	0.1595
10	0.0289	0.1591	0.1769
15	-0.0082	0.2247	0.2027
20	0.0579	0.2941	0.2340
		$\sigma^2=100$ $\alpha=10$	
0	-0.0225	0.3391	0.3391
2	-0.0216	0.3440	0.3397
3	-0.0183	0.3397	0.3404
4	-0.0265	0.3434	0.3414
5	-0.0205	0.3413	0.3426
10	-0.0204	0.3613	0.3530
15	-0.0233	0.3521	0.3695
20	-0.0331	0.4005	0.3916
		$\sigma^2=100$ $\alpha=40$	
0	-0.0266	0.3680	0.3680
2	-0.0231	0.3678	0.3686
3	-0.0290	0.3719	0.3694
4	-0.0289	0.3681	0.3704
5	-0.0267	0.3622	0.3718
10	-0.0289	0.3744	0.3830
15	-0.0010	0.3991	0.4010
20	-0.0614	0.4237	0.4249

[1] (S.E.) for ungrouped data, i.e., for h=0

Table 4: Mean and S.E. of ungrouped and mid-point estimates of α, obtained in the simulation-model (i), $n_k=100$

h	Mean	S.E.	$(S.E.)^1\sqrt{(1+h^2/(12\sigma^2))}$
	$\sigma^2=25$ $\alpha=10$		
0	6.9964	0.3066	0.3066
2	9.9933	0.3047	0.3086
3	10.0006	0.3088	0.3112
4	9.9916	0.3068	0.3147
5	9.9885	0.3258	0.3191
10	10.0075	0.3706	0.3540
15	9.5020	0.3555	0.4056
20	9.9869	0.2672	0.4683
	$\sigma^2=25$ $\alpha=40$		
0	40.0149	0.2892	0.2892
2	40.0157	0.2916	0.2911
3	40.0187	0.2934	0.2935
4	40.0156	0.2846	0.2968
5	40.0263	0.3044	0.3010
10	40.0323	0.3227	0.3339
15	39.6091	0.3658	0.3826
20	40.0225	0.5825	0.4418
	$\sigma^2=100$ $\alpha=10$		
0	9.9118	0.6253	0.6253
2	9.9156	0.6259	0.6263
3	9.9168	0.6239	0.6276
4	9.9184	0.6359	0.6294
5	9.9153	0.6302	0.6318
10	9.9069	0.6501	0.6508
15	9.9125	0.7256	0.6814
20	9.9275	0.6911	0.7220
	$\sigma^2=100$ $\alpha=40$		
0	39.9574	0.5485	0.5485
2	39.9616	0.5509	0.5494
3	39.9559	0.5485	0.5505
4	39.9684	0.5648	0.5521
5	39.9627	0.5649	0.5541
10	39.9619	0.5850	0.5709
15	39.9344	0.6133	0.5977
20	39.9755	0.6747	0.6333

[1] (S.E.) for ungrouped data, i.e., for h=0

Table 5: Mean and S.E. of ungrouped and mid-point estimates of β, obtained in the simulation-model (i), $n_k=100$

h	Mean	S.E.	(S.E.)[1]$\sqrt{(1+h^2/(12\sigma^2))}$
	$\sigma^2=25$	$\alpha=10$	
0	-0.0027	0.0530	0.0530
2	-0.0019	0.0530	0.0533
3	-0.0038	0.0537	0.0538
4	-0.0012	0.0540	0.0544
5	-0.0019	0.0573	0.0551
10	-0.0036	0.0644	0.0612
15	0.0024	0.0660	0.0701
20	0.0006	0.0484	0.0809
	$\sigma^2=25$	$\alpha=40$	
0	-0.0034	0.0563	0.0563
2	-0.0038	0.0567	0.0567
3	-0.0043	0.0559	0.0591
4	-0.0048	0.0554	0.0578
5	-0.0044	0.0598	0.0586
10	-0.0052	0.0619	0.0650
15	-0.0122	0.0703	0.0748
20	-0.0042	0.1082	0.8598
	$\sigma^2=100$	$\alpha=10$	
0	0.0183	0.1241	0.1241
2	0.0177	0.1237	0.1243
3	0.0186	0.1225	0.1245
4	0.0172	0.1249	0.1249
5	0.0178	0.1256	0.1253
10	0.0169	0.1309	0.1291
15	0.0177	0.1385	0.1352
20	0.0114	0.1347	0.1432
	$\sigma^2=100$	$\alpha=40$	
0	-0.0013	0.1058	0.1058
2	-0.0018	0.1063	0.1060
3	-0.0001	0.1046	0.1062
4	-0.0034	0.1080	0.1065
5	-0.0026	0.1082	0.1069
10	-0.0009	0.1122	0.1101
15	0.0019	0.1162	0.1153
20	0.0023	0.1275	0.1221

[1] (S.E.) for ungrouped data, i.e., for h=0

Table 6: Mean and S.E. of ungrouped and mid-point estimates of β_1, β_2 and β_3—model (ii), σ=100

h	Mean		
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
		n_k=10 for all k	
0	0.7193	-1.5555	0.0191
10	0.7110	-1.4928	-0.0219
50	0.6257	-1.2162	-0.1175
100	0.6832	-0.7006	-0.5402
200	0.2899	-0.8026	-0.2323
		n_k=25 for all k	
0	-0.3682	2.2355	-0.9839
10	-0.3674	2.2426	-0.9911
50	-0.3822	2.4762	-1.1478
100	-0.3151	2.4754	-1.2884
200	-0.2097	1.5684	-0.7932
		n_k=100 for all k	
0	0.1319	-0.2491	0.0450
10	0.1283	-0.2353	0.0370
50	0.1679	-0.2946	0.0333
100	0.2409	-0.3710	0.0321
200	0.2629	-0.4829	0.0933

Table 7: Covariance matrix of the ungrouped and the mid-point estimates of β and the values of $(\underset{\sim}{x}'\underset{\sim}{x})^{-1}\hat{\sigma}^2(1+h^2/(12\ \tilde{\sigma}^2))$ in parentheses for comparison-model (ii), n_k=10, σ=100.

	$\hat\beta_1$	$\hat\beta_2$	$\hat\beta_3$
		Ungrouped data	
$\hat\beta_1$	60.84	-183.59	57.29
$\hat\beta_2$	-183.59	803.98	-326.87
$\hat\beta_3$	57.29	-326.87	156.92
		Groups of length h=10	
$\hat\beta_1$	60.67(60.89)	-183.16(-183.75)	57.16(57.34)
$\hat\beta_2$	-183.16(-183.75)	802.72(804.65)	-326.30(-327.14)
$\hat\beta_3$	57.16(57.34)	-326.30(-327.14)	156.57(157.05)
		Groups of length h=50	
$\hat\beta_1$	62.49(62.10)	-187.20(-187.39)	57.50(58.48)
$\hat\beta_2$	-187.20(-187.39)	823.52(820.63)	-333.86(-333.63)
$\hat\beta_3$	57.50(58.48)	-333.86(-333.63)	160.98(160.17)
		Groups of length h=100	
$\hat\beta_1$	64.30(65.88)	-191.02(-198.80)	58.45(62.04)
$\hat\beta_2$	-191.02(-198.80)	842.92(870.57)	-341.83(-353.94)
$\hat\beta_3$	58.45(62.04)	-341.83(-353.94)	165.69(169.91)
		Groups of length h=200	
$\hat\beta_1$	77.20(81.00)	-233.75(-244.42)	70.13(76.27)
$\hat\beta_2$	-233.75(-244.42)	1061.75(1070.33)	-435.38(-435.15)
$\hat\beta_3$	70.13(76.27)	-435.38(-435.15)	217.81(208.90)

Table 8: Covariance matrix of the ungrouped and the mid-point estimates of $\underset{\sim}{\beta}$ and the values of $(\underset{\sim}{x}'\underset{\sim}{x})^{-1}\hat{\sigma}^2(1+h^2/(12\sigma^2))$ in parentheses for comparison-model (ii), $n_k=25$, $\sigma=100$

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
		Ungrouped data	
$\hat{\beta}_1$	17.82	-57.17	20.72
$\hat{\beta}_2$	-57.17	257.42	-111.59
$\hat{\beta}_3$	20.72	-111.59	55.42
		Groups of length h=10	
$\hat{\beta}_1$	17.69(17.83)	-56.81(-57.22)	20.68(20.74)
$\hat{\beta}_2$	-56.81(-57.22)	256.56(257.63)	-111.53(-111.68)
$\hat{\beta}_3$	20.68(20.74)	-111.53(-111.68)	55.42(55.47)
		Groups of length h=50	
$\hat{\beta}_1$	17.12(18.19)	-55.23(-58.36)	20.17(21.15)
$\hat{\beta}_2$	-55.23(-58.36)	255.82(262.78)	-112.54(-113.91)
$\hat{\beta}_3$	20.17(21.15)	-112.54(-113.91)	56.56(56.57)
		Groups of length h=100	
$\hat{\beta}_1$	18.03(19.30)	-56.92(-61.93)	20.55(22.45)
$\hat{\beta}_2$	-56.92(-61.93)	263.66(278.87)	-116.89(-120.89)
$\hat{\beta}_3$	20.55(22.45)	-116.89(-120.89)	59.58(60.04)
		Groups of length h=200	
$\hat{\beta}_1$	21.87(23.76)	-63.65(-76.23)	21.94(27.63)
$\hat{\beta}_2$	-63.65(-76.23)	307.61(343.23)	-39.40(-148.79)
$\hat{\beta}_3$	21.94(27.63)	-139.40(-148.79)	73.75(73.89)

Table 9: Covariance matrix of the ungrouped and the midpoint estimates of $\underset{\sim}{\beta}$ and the value of $(\underset{\sim}{x}'\underset{\sim}{x})^{-1}\hat{\sigma}^2(1+h^2/(12\ \sigma^2))$ in parentheses for comparison—model (ii), $n_k=100$, $\sigma=100$

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
		Ungrouped data	
$\hat{\beta}_1$	4.02	-11.94	3.67
$\hat{\beta}_2$	-11.94	58.05	-24.35
$\hat{\beta}_3$	3.67	-24.35	12.18
		Groups of length h=10	
$\hat{\beta}_1$	4.04(4.02)	-12.00(-11.95)	3.70(3.67)
$\hat{\beta}_2$	-12.00(-11.95)	58.38(58.10)	-24.50(-24.37)
$\hat{\beta}_3$	3.70(3.67)	-24.50(-24.37)	12.24(12.19)
		Groups of length h=50	
$\hat{\beta}_1$	4.10(4.10)	-12.20(-12.19)	3.83(3.75)
$\hat{\beta}_2$	-12.20(-12.19)	58.25(59.26)	-24.42(-24.86)
$\hat{\beta}_3$	3.83(3.75)	-24.42(-24.86)	12.15(12.43)
		Groups of length h=100	
$\hat{\beta}_1$	4.28(4.35)	-12.55(-12.93)	3.99(3.98)
$\hat{\beta}_2$	-12.55(-12.93)	61.15(62.89)	-26.21(-26.38)
$\hat{\beta}_3$	3.99(3.98)	-26.21(-26.38)	13.24(13.20)
		Groups of length h=200	
$\hat{\beta}_1$	4.79(5.36)	-14.14(-15.92)	4.47(4.89)
$\hat{\beta}_2$	-14.14(-15.92)	69.39(77.40)	-29.61(-32.47)
$\hat{\beta}_3$	4.47(4.89)	-29.61(-32.47)	15.18(16.24)

and family history of hypertension, blood and sugar levels of urine were also recorded. We study the dependence of systolic and diastolic blood pressure levels on age and sex, only for the sake of illustration. The paucity of observations on other variables does not allow extensive analysis of the data. Persons between ages of 18 and 28 were involved in the study and there were about 4,000 observations.

The bl~od pressure levels were grouped into two different intervals--intervals of size 5 mm Hg and 10 mm Hg. Since digit preference in recording and reading blood pressure measuring instruments is well documented, see for reference Borhani et. al. (1969), in view of this, the size of interval equal to 10 mm Hg seems quite reasonable and the error will be eliminated if we choose the size equal to 10. Table 10 provides the estimates with the standard errors for the parameters of the model for systolic and diastolic pressure given by the equations.

$$\text{Systolic Pressure} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Sex}) + \text{error}$$

$$\text{Diastolic Pressure} = \gamma_0 + \gamma_1(\text{Age}) + \gamma_2(\text{Sex}) + \text{error}$$

Table 10: Estimates of the Parameters (Estimate ± Standard Error)

Parameter	Systolic Pressure		
	h = 0	h = 5	h = 10
β_0	111.71 ± 1.54	112.72 ± 1.55	113.77 ± 1.58
β_1	0.05 ± 0.07	0.04 ± 0.07	0.01 ± 0.08
β_2	11.54 ± 0.39	11.61 ± 0.39	11.69 ± 0.40
	Diastolic Pressure		
	h = 0	h = 5	h = 10
γ_0	62.71 ± 2.27	63.65 ± 2.28	64.95 ± 2.31
γ_1	0.47 ± 0.11	0.47 ± 0.11	0.43 ± 0.11
γ_2	4.84 ± 0.57	4.88 ± 0.57	4.99 ± 0.58

ACKNOWLEDGEMENTS

We are grateful to Dr. Walton R. Garner of the Student Health Services, and to Dr. Richard R. Lanese of the Department of Preventive Medicine, The Ohio State University, for their permission to use the blood pressure data. Comments on an earlier draft of the paper by Professor Keith Eberhardt are highly appreciated.

REFERENCES

Borhani, N. O., Slansky, O., Gaffey, W., and Borkman, T. (1969). Familial aggregation of blood pressure. American Journal of Epidemiology, 89, 543.

Durbin J. (1954). Errors in variables. Review of the International Statistical Institute, 22, 23-52.

Fryer , J. G., and Pethybridge, R. J. (1972). Maximum likelihood estimation of a linear regression function with grouped data. Applied Statistics, 21, 142-154.

Gjeddeback, N. F. (1949). Contribution to the study of grouped observations: Application of the method of maximum likelihood in case of normally distributed observations. Skandinavisk Aktuarietidskrift, 32, 135-139.

Haitovsky, Y. (1973). Regression Estimation from Grouped Observations, Hafner Press, New York, N. Y.

Kendall, M. G., and Stuart, A. (1969). The Advanced Theory of Statistics, Vol. I, Hafner Publishing Co., New York, N. Y.

Kulldorf, G. (1961). Estimation from Grouped and Partially grouped samples, John Wiley and Sons, New York.

Tallis, G. M. (1967). Approximate maximum likelihood estimation from grouped data. Technometrics, 9, 599-606.

Thompson, W. A., Jr. (1977). On the treatment of grouped observations in life studies. Biometrics, 33, 463-470.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR No. 166	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) "Approximate Maximum Likelihood Estimates in Regression Models for Grouped Data"		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) A. Indrayan J. S. Rustagi		8. CONTRACT OR GRANT NUMBER(s) NR 042-403 N00014-78-C-0543
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Ohio State University Columbus, Ohio		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE 1978
		13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Statistics and Probability Program, Code 436 Arlington, VA 22217		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)